Information Discovery and Retrieval Tools

Michael T. Frame
U.S. Geological Survey, Center for Biological Informatics
Mail Stop No. 302, 12201 Sunrise Valley Drive
Reston, Virginia 22092
USA

mike_frame@usgs.gov

ABSTRACT

Due to the rapid growth of electronically accessible content from the Internet, there is a corresponding increase in demand for information of all types from a number of diverse users. Although the World-Wide Web presents tremendous opportunities to users for access to this wealth of information, the quantity of that information can be overwhelming. The user who attempts to find information can become confounded by the sheer volume of data and information returned as "pertinent" to his/her need. In addition, current awareness becomes an obstacle, as variations in search engine crawls of the Web, as well as the user's own ability to keep up with frequent queries to multiple search tools, can prevent timely access to and knowledge of pertinent information. This session will focus on the various Internet search engines, directories, and how to improve the user experience through the use of such techniques as metadata, meta-search engines, subject specific search tools, and other developing technologies.

1.0 BACKGROUND

Ever since the Internet's beginnings in the 1990s, the amount of information available on the World-Wide Web has steadily increased. It is estimated that close to 10 billion web pages exist on the World-Wide Web today. As expected, this number is continuing to grow; however, at a much slower and some say more controlled rate. The rate of growth of World-Wide Web content has also caused the community of casual and advanced users, to consider alternative means to finding information.

As the information content has grown on the World-Wide Web, so too has the need for improved tools and products to aid users in this discovery of information. Several tools basically perform the same function, but may differ slightly in their methods and results. This primarily has to do with vendor specific interpretation of World-Wide Web terms such as: Spam, spider/crawler configurations, and collection size. All of this leads to industry estimates that less than 20% of the entire content of the World-Wide Web is available to the typical user (World-Wide Web Consortium 2004). This paper investigates various terminologies and provides simple techniques users can perform to improve their search experiences on the World-Wide Web.

2.0 BASIC TERMINOLOGY

2.1 What Do Internet Search Engines Really See?

From a user's perspective, as shown in Figure 1, users often simply enter a term in a simple search box and wait for results. They are oblivious to what the computer or system is doing. This is the way it should be. If users have to worry about how an Internet search engine is configured or what it expects, then most likely

Paper presented at the RTO IMC Lecture Series on "Electronic Information Management", held in Sofia, Bulgaria, 8-10 September 2004, and published in RTO-EN-IMC-002.

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
00 DEC 2004	**N/A**	**-**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Information Discovery and Retrieval Tools	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **U.S. Geological Survey, Center for Biological Informatics Mail Stop No. 302, 12201 Sunrise Valley Drive Reston, Virginia 22092 USA**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release, distribution unlimited

13. SUPPLEMENTARY NOTES
See also ADM001735, RTO-EN-IMC-002, Electronic Information Management., The original document contains color images.

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**	**UU**	**8**	

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std Z39-18

the search engine user interface needs to be redesigned or another product selected. Users have too many other things to do, whether at work or home, to concern themselves with learning the various idiosyncrasies of each Internet search engine.

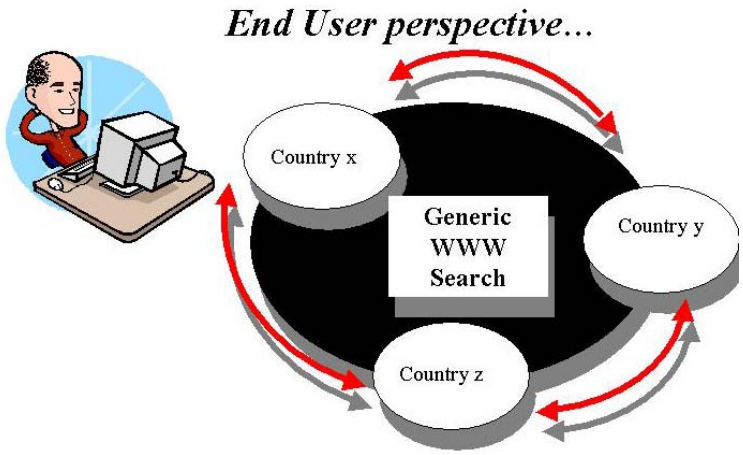

Figure1: Typical User Search.

However, what the user often does not realize is that Internet search engines primarily read the underlying document codes or "metatags" within a document. Metatags are document tags or properties that are often stored within the Header of an HTML document or within the document itself. Figure 2 below describes a typical view that an Internet search engine would see when it indexes a document.

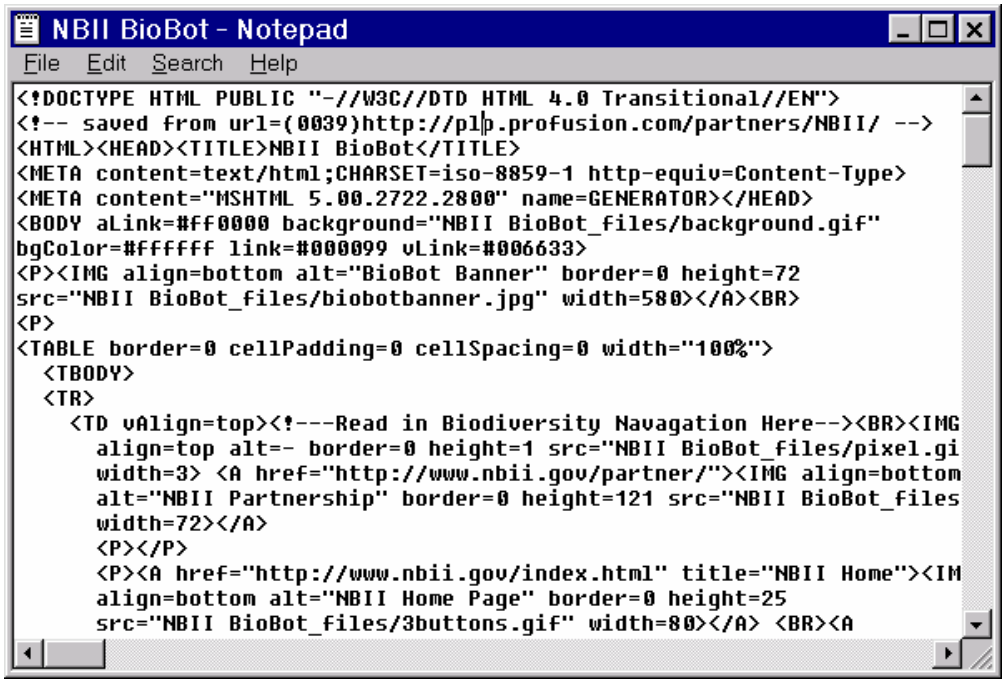

Figure 2: Typical Internet Document as Viewed by Search Engines.

2.2 What is Spam?

"Spam" is a term you often hear thrown about on the World-Wide Web today. Spam is not just a popular Hawaiian luncheon meat anymore. Understanding what spam is and is not is very important in understanding how search engines on the WWW discover and display information to users. Spam is considered to be anything that a software developer or HTML creator does to try to falsify his or her content to a web engine. In today's web environment content creators jockey for position on Internet search engines results/hits lists and often resort to categorizing their sites in ways that may not truly represent the content or overall purpose. This is considered spamming a search engine crawler or data harvester. Tricks commonly employed by web content creators include applying keywords within the Header section of an HTML document that have nothing to do with their site, or simply creating BLANK HTML pages with white text so that users don't see the content, but a search engine can. Internet Search Engines are all wise to these tricks and this is why it is often difficult for content producers and/or developers who have truthful content and are trying to do a good job in making their content available understand what an Internet search engine expects and applies preferences to.

2.3 The Basic Internet Search Engine Model

Internet search engines on the WWW "harvest" data from publicly available web sites via automated jobs or crawls. This harvesting or gathering of summary information (usually items such as URL, keywords, summary description) to a central point is done with spiders and/or crawlers. Spiders and crawlers are simply automated jobs or processes that run from an Internet search engine provider's server and scour the WWW for content. This content is then made available through the Internet search engine providers' central index. Figure 3 below demonstrates this process.

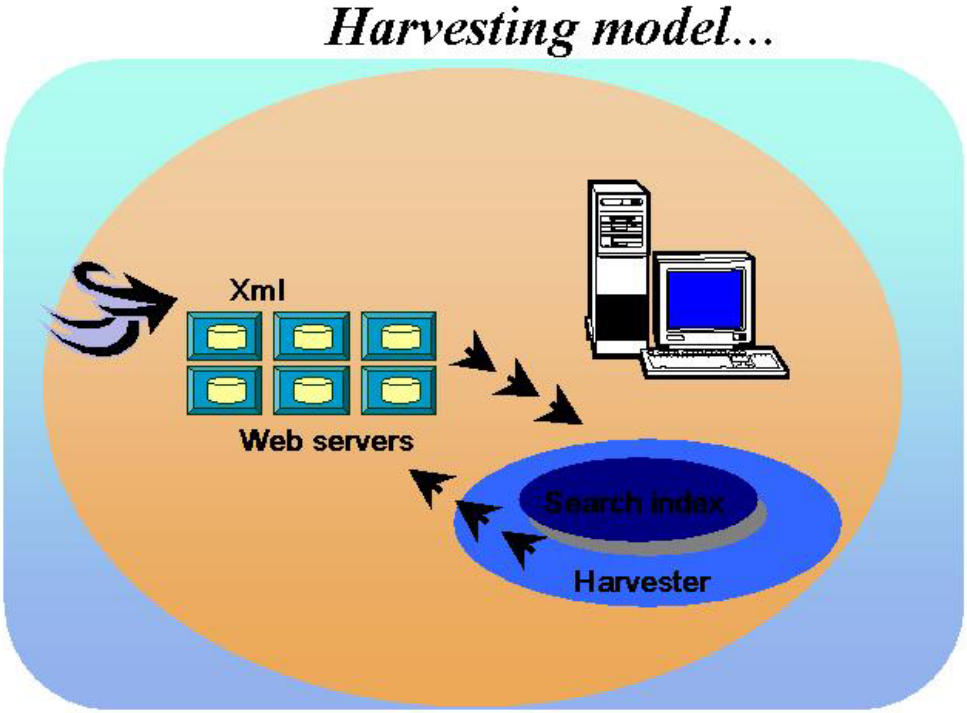

Figure 3: Basic Internet Search Engine Harvesting Model.

2.4 What are Metatags and Why are they Important?

Embedding metatags within the HTML of your Web site not only promotes higher rankings, and thus, better retrieval, of your site by many of the major search engines, but also provides a foundation for future information retrieval and discovery on the Web as the web evolves into a more structured organization of content. The algorithms used by search engines constantly change; however, the presence of metatags on your pages can often make a dramatic difference in enabling users to find your information. Remember, too, that as various sites apply metatags, an **integrated** system whereby users can easily locate your site through a search engine are likely to explore other related sites within the WWW.

The table below describes both standard metatags and unique discipline, in this example biological information, metatags that all can be implemented on web sites. Search engines require some tags, while others are optional, depending upon the scope and context of the page(s) under development. Additional metatag requirements may be added as retrieval tools become more sophisticated. Fortunately, the creation and editing of metatags is a quick and simple process, thanks to the development of metatag software, which can rapidly generate tags selected by a content provider across designated pages, directories, or an entire site.

The metatags in Table 1 below are all standard HTML 3.0 or above supported tags. If users are using dynamically created web sites, the metatags described below can simply be created automatically out of a database dump or export.

Table 1: Recommended Metatags

Metatag	Definition	Format & Sample Value
Author	The Author Tag contains name of the content provider (*not* the Webmaster / programmer).	<meta name="author" content="Bob Johnson">
Title	Even though the Title tag is not considered a true metatag, it is critical in search engines' ranking algorithms, and provides users with general information about your page. Search Engines results/hit lists also display the Title tag. Up to 80 characters can be contained within this tag.	<TITLE> West Nile Virus: Wildlife Impacts - NBII</TITLE> *** please maintain this format when naming your pages ***
Keywords	Keywords are probably the most important meta-tag that a Web site manager can include. Up to 1000 characters can be contained within this tag. *Your keyword contents should include the basic tags at left, plus all terms relevant to your site and particular sub-sections. Include several generic terms that apply to your entire node, plus terms specific to various sub-directories and pages. Try to think of as many synonyms for your terms as you can. Note that you need to include term variations (e.g. bird, birds, birding, birdwatcher), as the search engines do not employ stemming when parsing keywords. Spelling counts! Use terms found within the page contents to boost relevancy rankings.*	<meta name="keywords" content="*your page-specific keywords….,* NBII, National Biological Information Infrastructure, biology, biodiversity, natural resources, reference, education, ….. ">> *place these standard keywords AFTER your page-specific keywords*
Page Description	The Description tag is used by search engines to display information about your page and to index its contents. Up to 200 characters can be contained within this tag. The description often determines whether the searcher will choose to view your page. Make the description relevant to the particular sub-section or page; don't rely on one generic description for all pages on your site. Use keyword tag terms in your description to boost term relevancy rankings.	<meta name="description" content="This is the textual description for your page. Please make sure your spelling is correct and include any relevant keywords within the Description tag.">
Language	Even though most content on the web is in English, the Language tag adds value to your Web site, helping users limit search engine retrieval to a particular language.	<meta name="language" content="en-us">
Classification	The Classification tag is often used by a number of the Web search engines when you register your site and/or when your site is indexed so that your site can be classified with other similar sites. Typical values include: "Government, Science, Education, etc."	<meta name="classification" content="Government, Science">
Ratings/PICS	The Ratings and PICS tags are used by Internet providers and search engines to limit access to a particular page. Often this is used to restrict access to "Mature Audience Only" pages for children using the Internet. Typical Values include: "General, Restricted, Mature, Safe for Kids", etc. Because filters are becoming more common within retrieval tools and browsers, or as added software, these tools may arbitrarily block your site if the tag is not implemented.	<meta name="rating" content="General, Safe for Kids">

Table 2 below describes the unique or custom metatags for a domain specific organization. In this case, these custom metatags are relevant to categorizing, displaying, and delivering biological data and information.

Table 2: Domain Specific Metatags (Custom Tags)

Metatag	Definition	Format & Sample Value
Species Scientific Name	The Scientific Name of a particular Species on the web page being classified. NBII Partners are strongly encouraged to utilize the Integrated Taxonomic Information System (ITIS) (http://www.itis.usda.gov/plantproj/itis/index.html) as its basis for completing this information.	<meta name="Species Scientific Name" content="Parnassius smintheus">
Species Common Name	The Common Name of a particular Species on the Web page being classified. The Common Name is extremely important to both expert and novice users for finding information about a particular species. ITIS is a source for completing this meta-tag.	<meta name="Species Common Name" content="Rocky Mountain Parnassian">
Organization	The lead Partner organization that maintains the specific Web site/page being classified. The use of standard controlled lists is strongly encouraged for completing this field.	<meta name="Organization" content="USGS Center for Biological Informatics">
Web-site Theme	The high-level Theme (Education, etc.) that your Web page falls under within a web structure.	<meta name="website Theme" content="Education">
Web–site Category	The specific Category, within the website Theme, that your Web page falls under.	<meta name="website Category" content="General Curriculum">

Domain specific metatags greatly aid a particular community of users in the discovery and identification of quality resources. For example, if a user accesses one of the search engines on the World-Wide Web today and searches for a specific bird, i.e. "common loon", the search result produces a hit list of more than 13 million results. Some of these results are most likely pertinent to the user, but most are not and it is infeasible for a user to navigate through 13 million web pages for relevant data.

To resolve this issue, programs such as the National Biological Information Infrastructure (http://www.nbii.gov) have been implementing a refined and improved spidering methodology with its partners and applying metatags within its local and partner pages. As a result, users can now easily narrow their results lists to 62,000 web pages with the same search that yielded over 13 million results. These spidered and indexed pages are primarily biological in nature and due to the intellectual effort that is currently ongoing within the NBII Program for adding information content to the NBII System, users can expect to receive more targeted and a higher quality result than directly access the WWW and its search engines. Users also have the ability to narrow their search results to 1,400 web pages and information sources through the direct querying of meta-information contained within a domain specific or custom meta-tag called "Common Name". As one can imagine, this saves users tremendous time and presents authoritative and related information to a user without requiring an already information overloaded user to review a large number of primarily non-pertinent results.

3.0 TYPICAL SEARCH ENGINE FEATURES AND CAPABILITIES

As stated, all search engines are mostly the same, but often different in their implementation and configurations. Below are some of the features you would expect to find in a typical search engine. Often low-end search engines may or may not have all of the features noted or may be limited in how many documents one may index or limited on the size of your collection.

- Contains an automated spider or crawler
- No theoretical limits in the amount of indexing (limited by hardware)
- Supports remote indexing
- Continual background indexing of content
- Custom metatag support (some low-end products do not support this feature)
- Support for indexing PDF, .doc, etc. (some low-end products do not support this feature)
- Supports URL and word exclusions & inclusions
- SSI supported
- Search by custom metatags
- Case sensitive or insensitive searching
- Simple search interface
- Ability to customize search results pages
- Boolean Searching capabilities
- Provide users meta description and page title in search results
- Inexpensive cost, – $200
- Easily customizable search/results interface
- Result weighting feature
- URL Inclusion list for target indexing
- Require significant memory (RAM) and disk space as the collection grows
- Low-end alternatives often do not possess the capabilities to do phrase or natural language searching.

4.0 WHAT CAN YOU DO AS A CONTENT DEVELOPER OR SOFTWARE DEVELOPER TO IMPROVE DISCOVERY OF YOUR CONTENT?

Users can do several things to help ensure that their information content is more readily found on the WWW today. Some of these things make perfect sense, but users often do not dedicate the necessary resources required to make them happen on a regular basis. Each environment and web site is different; however, the general principles and techniques noted below will help any web content producer.

- Implement metatags on your and your partners web sites
- Update content frequently
- Register your site with the major search engines (tools exist to aid in this process)
- Perform a basic study of where your site results within the major search engine providers

- Do not spam the search engine providers
- Re-evaluate your web site directory structure to ensure information is appropriately categorized/ described within your URL strings
- Look through your server log files to determine what users are trying to find on your site and/or the path they are using to find information
- Perform basic usability testing of your site to determine what users expect and can easily gather from your site. This also may determine why users go to an Internet search engine provider versus accessing your site directly.
- Realize that Internet search engines don't all act the same, index at the same time period, and often value a particular metatag, document date, etc. more than another vendor product.

5.0 CONCLUSION

As one can see, maintaining awareness and improving delivery of your information via the WWW in today's environment is almost a full-time job. As Internet search engine providers become more sophisticated, so too will it be necessary for content producers and providers to restructure their information to take advantage of such capabilities. With the advent of new technologies, such as XML and SOAP, information content will be more readily able to be delivered at a more granular scale and to a more targeted audience. However, these technologies are still in their infancy, as it comes to the overall web content, and Internet search engines will continue to be one of the major sources whereby users access to gather information.

6.0 REFERENCES AND FURTHER READING

Arasu, A., Cho, J., Garcia-Molina, H., Paepcke, A., and Ragahavan, S. (2001). Searching the web. *ACM Transactions on Internet Technology* 1(1): 2-43.

Kobayashi, M. and Takeda, K. (2000). Information retrieval on the web. *ACM Computing Surveys* 32(2): 144-73.

Nobles, R. and O'Neill, S. (2000). *Streetwise maximizing web site traffic: build web site traffic fast and free by optimizing search engine placement.* Avon, MA: Adams Media Corporation.

Notess, G. (2004). *Search Engine Showdown Homepage*. [Online]. Available: http://www.searchengineshowdown.com [5 July 2004].

President's Committee of Advisers on Science and Technology (PCAST) Panel on Biodiversity and Ecosystems. (1998). *Teaming with life: investing in science to understand and use America's living resources.* [Online]. Available: http://www.nbii.gov/about/pubs/twl.pdf [9 May 2002].

Sullivan, D. (2004). *Search Engine Watch Homepage*. [Online]. Available: http://searchenginewatch.com [5 July 2004].

World-Wide Web Consortium Homepage. (2004). [Online]. Available: http://www.w3c.org [5 July 2004].

Yonaitis, R. (2000). *The elements of <web site> promotion.* Concord, NH: Hiawatha Island Software Corporation.